Alberto Bordino and Olga Klopp's contribution to the Discussion of "Statistical exploration of the Manifold Hypothesis" by Whiteley et al.

Alberto Bordino[1] and Olga Klopp[2]

[1]Department of Statistics, University of Warwick
[2]ESSEC Business School

Address for correspondence: Alberto Bordino, Department of Statistics, Mathematical Sciences Building, University of Warwick, Gibbet Hill Road, Coventry, CV47AL, UK. *Email:* alberto.bordino@warwick.ac.uk

We congratulate the authors on an insightful, practice-grounded contribution showing how manifold-like structure can arise naturally from a broad latent metric space (LMS) model. We especially commend the "practice-first" perspective: beginning with pipelines common among practitioners (e.g., PCA to moderate dimension, then t-SNE/UMAP), and then developing theory that explains why such workflows succeed.

We shall now comment on a point where practitioners may require further guidance. In particular, Section 4.3 rightly recommends normalising PCA scores when per-sample amplitudes dominate, since spherical projection preserves directions while removing gains. But if nuisance variation is anisotropic, spherical projection can distort neighbourhoods. A "whiten-then-sphere" variant, with stability checks, can clarify geometry. To illustrate this point, we show that an elliptical transformation can be appropriate in certain settings, and empirically validate this claim on the temperature-dataset.

As in the paper, we assume $f(z, z') = \langle \phi(z), \phi(z') \rangle$ and that the associated manifold $\mathcal{M}$ is a subset of the unit hypersphere. Now, fix an invertible $A \in \mathbb{R}^{r \times r}$ and i.i.d. $V_j \in \mathbb{R}^r$ with $\mathbb{E}[V_j] = 0$, $\mathbb{E}[V_j V_j^\top] = I_r$, and finite fourth moments, and define

$$X_j^{(A)}(z) := V_j^\top A\, \phi(z), \qquad Y_{ij} = \alpha_i X_j^{(A)}(Z_i) + \sigma E_{ij},$$

with $\{\alpha_i\}_{i=1}^n$ i.i.d. positive random variables. In light of Proposition 1, this model serves as an extension of (21) and coincides with it when $A = I_r$. We have

$$f_A^{\mathrm{ext}}(\alpha, z, \alpha', z') := \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}[\alpha X_j^{(A)}(z)\, \alpha' X_j^{(A)}(z')] = \langle A\{\alpha \phi(z)\}, A\{\alpha' \phi(z')\} \rangle =: \langle \phi_A^{\mathrm{ext}}(\alpha, z), \phi_A^{\mathrm{ext}}(\alpha', z') \rangle,$$

so that applying Theorem 1 in this setting yields $\|p^{-1/2} Q \zeta_i - \phi_A^{\mathrm{ext}}(\alpha_i, Z_i)\| \xrightarrow{\mathbb{P}} 0$. It is immediate to see that a spherical projection removes α_i but not A, hence PCA would recover the A-warped manifold $M_A := \{A\, \phi(z) :$

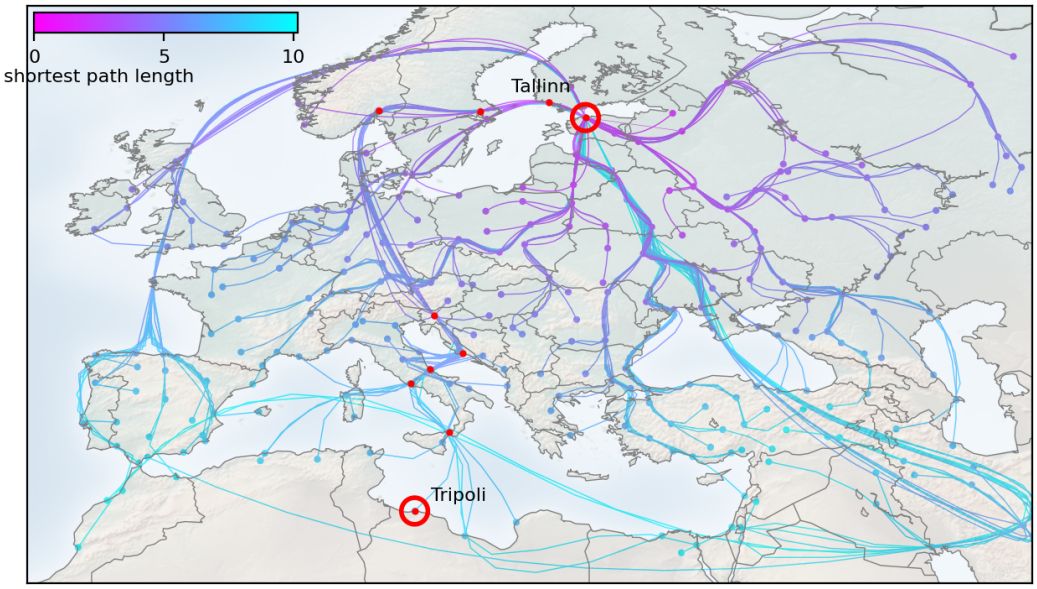

Figure 1: Analogue of Figure 12 in the paper, where PCA scores undergo the elliptical transformation $\zeta \mapsto \widehat{\Sigma}^{-1/2}\zeta$ before being projected onto the sphere. Here, the shortest path in the graph embedding from Tallin to Tripoli is closer to the geographically shortest path. Code to reproduce the simulation can be found at https://github.com/abordino/ExploreManifoldHypothesis.

$z \in Z\}$ up to an unknown rotation. To recover $\mathcal{M}$, we propose to apply an elliptical transformation based on $\widehat{\Sigma} := n^{-1}\sum_{i=1}^{n}\zeta_i\zeta_i^\top$ before the final spherical step. Indeed, if the latent distribution is approximately isotropic, i.e. $\mathrm{Cov}\{\phi(Z)\} = c\,I_r$, then heuristically $\widehat{\Sigma} \approx \mathbb{E}[\zeta_1\zeta_1^\top] \approx c\,p\,\mathbb{E}[\alpha_1^2]\,Q^\top AA^\top Q$ as $r \ll n$, and

$$p^{-1/2}Q\,\widehat{\Sigma}^{-1/2}\zeta_i \approx \frac{\alpha_i}{\sqrt{cp\,\mathbb{E}[\alpha_1^2]}}\,(AA^\top)^{-1/2}A\,\phi(Z_i),$$

where $(AA^\top)^{-1/2}A$ is orthogonal. Hence whitening undoes A up to rotation, and the subsequent spherical projection removes the prefactor depending on α_i. The simulation results in Figure 1 demonstrate the pipeline's effectiveness for the temperature-data study.

We conclude with some questions for the authors:

1. How sensitive would such a procedure be to departures from the isotropy heuristic, e.g. $\mathrm{Cov}\{\phi(Z)\} = \Sigma_0 \neq cI_r$?

2. Can Theorem 1 be extended to produce a bound that that cleanly separates (i) the PCA approximation term (as in (16)), (ii) the covariance-estimation term for $\widehat{\Sigma}^{-1/2}$, and (iii) a model-misspecification term reflecting $\Sigma_0 \neq cI_r$?